Subject: Date:

text analysis :- 1 lec 14

* main Challenges.

1 High - dimensionality. as every distinct term is a dimension, when analyzing a document every possible word represent a dimension

2 Data is unstructured. 3 think about problem the right way

* Text Analysis Problem Solving Tasks (steps)

1 **parsing**. "impose structure on unstructured or semi structured text

2 Search / Retrieval

هل بندور على كلمة أو topic معين ولا الdocuments كلها

→ which documents have this word?

→ which ,,,, are about this topic or entity?

3 Text mining :- هنا بيصنف الdocments من معلوماتها

A → understand the document

B → classification or Clustring

Labe لو معاك → (classification)

تصنيف تحت بعض

لو معاكش label → (Clustring)

Subject: ______ [2] Date: ______

* Tasks are not ordered list :
يعني مش لازم تمشي بترتيب خطوات معينة
على حسب الاحتياج

* Usually you start with parsing then do either search or text mining

---

تطبيق عملي على الـ text :-

لو شركة كبيرة عايزة تعرف رأي الناس عن منتجاتها مثلاً
ولو الناس عاجباها (الإيجابيات +ve والسلبيات -ve

Steps :- " ودي بتشتغل إن كل مرحلة بالـ technique المعينة عشان نعمل "

① Monitor social network, review sites mention of our products.
وهنا بنبدأ نشوف رأي الناس ع الواقع ونتزل الداتا دي و ونستعمل
[1] parse the data feeds to get actual content
[2] find and filter the raw text to productname (use Regular Expression)

---

② Collect the review :-
① Extract the relevant raw text.
② Convert the raw text into suitable document representation
③ index our review corpus.

③ Sort the review by product

↳ classification

ودى معناه انك هتجمع الآراء المرتبطة بكل منتج لوحدها

4 Are they good reviews or bad?

↳ classification

⑤ Maketing calls up and reads selected reviews

↳ Search / information Retrievl

→ parsing the feeds.

① impose structure on semi-structured data

② We need to know where to look for what we're looking for.

→ regular Expression : finding words strings or particular patterns in text

→ match is Boolean response

يعنى انت مثلاً حاطط فالكود بتاعك كلمة بتدور عليها فوسط الـ document

هو رده يا اما الكود ده فيه حاجة شبه كده او لا لا

← يعنى رموز لما تكتب فالـ regular

b[P|p]hone	{|}	دى معناه أو	bPhone or bphone
bEbo*k	{*}	يعنى ممكن تكون متكررة	bEbook / bEbk / ---
^I love	{^}	(^) بيطبع أول الجملة دى	
Ahmed$	{$}	بيطبع آخر الجملة دى	

② Collect the reviews

يعني انت عارض علة محسنة اللي هي

I Love LOVE my bPhone!

هتعد من كل document كل كلمة اتكررت كام مرة

وبتعمل حاجة اسمها Bag of words

Bag of Words :. it's Common representation and vector with one dimension for every unique item

term frequency — (tf) حاجة اسمها

حدد بس عايزه تكرار الكلمات المهمة زي مثلا good, bad, love..

tf :- number times a term occurs

كلام مهم يعني كلام تقدر تستنتج منه رأي الناس فعلا ،

انما كلمة the مثلا مش مهمة مهما اتكررت .

acme	0
bPhone	1
Love	2
Slow	0
|	

طيب كل كلمة انت هتعطيها أو بتديها أولوية بتزود الـ dimensionality وده ممكن يعمل ال problem

نقوم عاملين تقليل ليها وهو إن نشيل الكلام اللي مش مهم أو له كلمة كان ليها معنى اتذكر قبل كده منحسبهاش

مثلاً

① stemming ② term space

phone = phones

no stop words : the , a

often no pronouns

Subject: ⑤ Date:

Collecting reviews وانت بتعمل

من هاتك جلا documents على اساس تعرف تشكل أسرع الـ مرتبة
بالوضوح اللى بتدور عليكو والا لا د ز ى

① features
Anything about the document that is used for search or analysis

② Title ③ tags or Keywords ④ Date info.
⑤ Source info ⑥ Named entities

---

3) Sort the reviews by product الـ text classification
بالراى اتطرقنا features

Judicious choice of features
① product mentioned in title
2) tweet, or review
3) Term frequency
④ Canonicalize abbreviations يعني الاختصار متعارف عليه
"5X" = "iPhone 5X"

---

④ Are they good reviews or bad?
① classification using Naive Bayes
بس محتاجين الاول نعمل الـ trian الـ algorithm من خلال
② tagged training data
وده مشكلة الـ text classification ازاي نجيب fit data for problem
① Hand-tagging ← وازاى الـ classification يتعمل ؟
② clues from review sites
③ cluster documents, then label the cluster

⑤ Marketing team - - -

① determine if the results you receive
are indeed the ones you want or not

② is this document what i wated?

③ used to rank seach result

precision:- what % of document relevat?

Recall .. of all relevent documents,
what % were returned to me?

يعني البيانات يلي عدد n من الـ document اللي ليها علاقة
بالموضوع انت قدرت تطلع كام واحد من الـ n.

---

* Computing relevance (Tf)

→ each term in a document a weight for that term.

→ weight of a term → t in document → d
is a function of the number
of times t appears in d

عدد مرات t in d

tf(t, d) = Count(t, d)

Inverse Document Frequency (Idf)

measure term uniqueness in corpus

indicates the importance of the term

for → search
→ classification

Subject: Date:

NLP(Natural language processing)

① unstructured text mining to extract features
its goal → vectorize the document

② after vectorization apply Machine learning
① clustring ② classification
Decision Tree
Naive Baysen

③ Scoringe :- categorize incoming documents